Analyzing Liquor Store Profitability in Iowa

Final Report for the Iowa Liquor Store Chain: Booze 'R' Us

Dylan Li, Alex Arrieta, Cameron Stivers, Ryuhei Shida, Nathan Hill

Introduction

Our goal with this project was to analyze the projected revenues of your company, Booze R Us, for the upcoming years. To form predictions about the fiscal success of Booze R Us we studied data collected by the Iowa Department of Revenue regarding the purchase of class E liquor sales by stores.[1] This data is an itemized list by single purchase and includes information about the purchasing store and its location, the brand and type of alcohol purchased, the quantity purchased, and the total sale amount in dollars. In addition to this information gathered from the Iowa state government, we also used the US census data to get accurate counts of the population in Iowa counties.[2] This data allowed us to predict the overall trends in alcohol sales in Iowa that should hold for Booze R Us' operations in the state.

[1] https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy

[2] https://www.iowa-demographics.com/counties_by_population

Data preparation

The initial set of alcohol purchase data contains a massive quantity of information with over 27 million recorded transactions. Thus, to gain a better understanding of the data we began by analyzing information from 2020 to 2022 when accessing individual purchase records. In addition to restricting the time frame of the data we were analyzing we also dropped either unnecessary or duplicate information. For example, sale quantity in gallons and sale quantity in liters describe the same metric, just in different units, thus only one needed to be kept. Information that was dropped due to lack of utility included items like a store's street address as this was too granular of a location to aggregate data by, and we had no way of inferring why particular addresses might outperform others. The data kept for the final analysis included: alcohol vendor and type, date, store county, per bottle and total price, and volume purchased. Appended to this data was population counts by county as that was the most granular location information we kept about a store.

When aggregating data across time the entire dataset from 2012 to 2022 was used due to the lower cost of working with aggregated data. We grouped the data by total sales per year and per month, in addition to sales over time per county and per alcohol type. This allowed us to observe multiple trend patterns over time and note if the more broad patterns observed were generally followed in smaller categories.

Model selection

We made several initial assumptions when approaching this problem. Most importantly it was assumed that Booze R Us revenue would change, on average across its stores, similarly to how alcohol sales in Iowa changed. Another critical assumption was that stores generally purchase as much as they sell. One key drawback here is the inability to tell how much a brand of alcohol was marked up, thus making true profit hard to discern. Still, stores will purchase alcohols that generate large profits, thus those alcohols should have larger total sales in the dataset.

The first step in designing the actual model was deciding which type of model specification to use, which based on the use of data primarily focused on a numeric response variable, a linear model specification was selected. Then to tune the parameters a multi-step process was used for each set of predictors considered. For every predictor first several loss functions were established which were Ordinary Least Squares (OLS) and Ordinary Least Squares with a Ridge Penalty with several different lambdas considered. Then for each loss function a 4-fold cross validation was done, with the average R Squared score across the 4 folds being stored. With the average R Squared for each loss function known, we simply elected to keep the loss function that maximized the R Squared. This model specification was then trained on the entire dataset to come up with a final model for a given set of predictors.

The most critical observation in the initial analysis of the dataset was the general stability in the increase of alcohol sales each year in the last decade which can be seen in Figure 1. Comparing the trends in total sales per year for the last decade and the trends in total sales per month for the last decade (Fig 2), despite the volatility in monthly sales, the yearly sales are nearly linear. Thus our first model parameters simply measured the average change of alcohol sales per year. For this model the optimal loss function was OLS which once the model was trained on the whole dataset resulted in an R^2 of 0.9516.

Another important business insight we decided to pursue investigating was potential store locations and so we also modeled how counties' alcohol purchases varied over time. Again here an OLS loss function was found to be optimal. This produced more varied results with the average R^2 being 0.7853 across all counties and 0.8511 being the median R^2. Some counties such as Iowa county were very unpredictable with an R^2 of only .012 (Fig 3). Additionally, Sioux and El Paso counties lacked enough data to model sales with.

One final model specification that was focused on was the trend in sales for different types of alcohol. Again an OLS loss function was more favorable than a loss function with a Ridge penalty. Several types of alcohol have sparse data and unreliable modeling results. This led to an average R^2 of 0.6316 and median R^2 of .7046 across all alcohol types with sufficient data that included up to 2022.

Final Model

The principal final model is the model examining sales trends per year in Iowa. Ultimately what justifies this model is the decade of sales that has proven to be predictable. This was by far the most accurate model we fit. The model shows that sales increase on average by 19 million dollars a year in Iowa as a whole. The loss of information about smaller time scale sales such as month was deemed unimportant due to the fact that the added volatility in monthly sales only served to obfuscate the overall trend in yearly sales. Additionally the overall success of individual stores is unlikely to be dependent on single monthly outcomes. Another key focus of this model is the aggregate data over trying to model individual stores. In deciding future expansion opportunities it was thought to be more important how Booze R Us would perform as a company than as a set of individual stores, and additionally the unpredictability in modeling any single store made it difficult to recommend store placements and stock at the individual level.

The county and alcohol type models act very similarly to the overall model for Iowa with a similar basis behind their justification. One thing worth noting is that these models were all constructed independently. This means there are 97 different models for counties, one for each county excluding Sioux and El Paso counties. This was done so that each individual model could be evaluated and tested systematically. However all 97 report the same type of results, and take in the same input, thus using all 97 in tandem is easily

achieved to get results for the whole state. In all cases what the coefficient of the year tells us is how much the county or alcohol type increases (or decreases) in total sales per year.

Key takeaway

The most important information gained was that Booze R Us can expect an about 4.5% increase in revenue in 2023, and another 4.3% increase in 2024. Again this assumes that Booze R Us sales will increase at the same rate that Iowa sales increase overall. This information should provide the necessary foresight in order to determine the extra capital Booze R Us will have to expand operations in the future. In addition to this revenue foresight, there are also clear implications about where to put new stores. The most lucrative counties by total sales are Polk, Linn, and Scott counties. However, the fastest growing markets are located in Fremont, Clinton, and Adams counties. It is also advisable to not add any stores to Iowa or Buchanan county as both have shrinking sales markets. For alcohol types vodka and tequila are incredibly popular alcohol types that are rapidly increasing in sales. Whiskies are popular but are not increasing in sales overall. Gins are incredibly unpopular and losing market share. These alcohol type recommendations can be used to tailor both new and existing store stocks. However, it should be noted that these numbers may vary by country which was not studied so some alcohol types may over or under perform from their average in certain counties.

Ethical Considerations

The data used in this process was publicly available data published by the State of Iowa. All usage of this data and model should follow Iowa State Government guidelines found here.[3] Additionally, despite this information being public, it should not be used against any competitors listed in the data directly or indirectly in manners such as attempting to slander these companies.

Another note is that these models are designed to simply inform business decisions by projecting possible revenue outcomes using several different indicators. We believe that you should still consider other possible ramifications of expanding your operations to certain locations such as promoting alcohol to vulnerable groups such as minors to avoid negatively impacting the lives of community members

[3] https://www.iowa.gov/policies

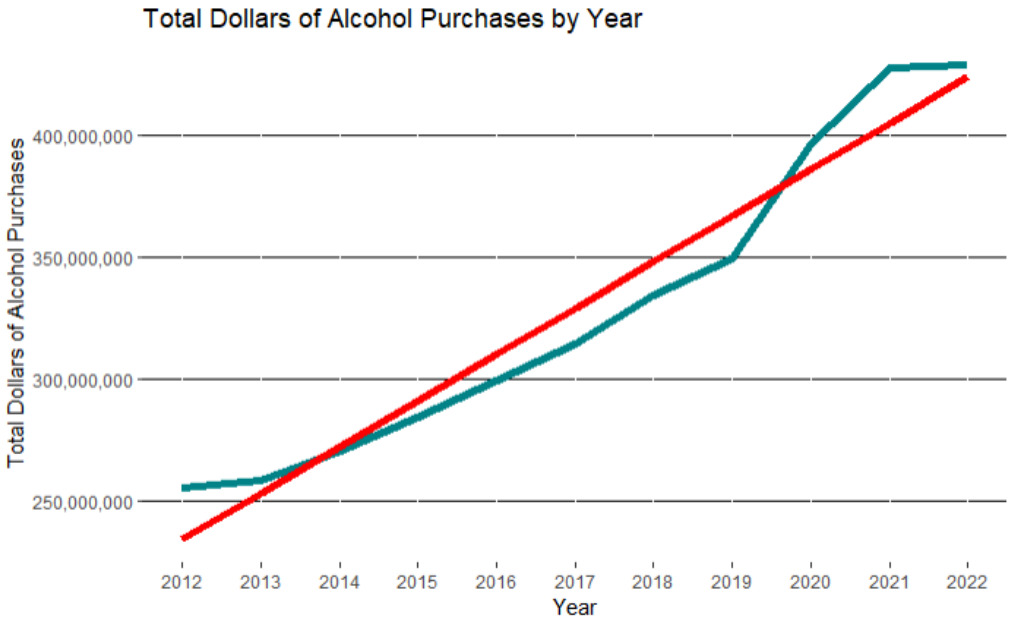

Figure 1: Year-to-year total alcohol purchase amounts in all of Iowa (dollars)

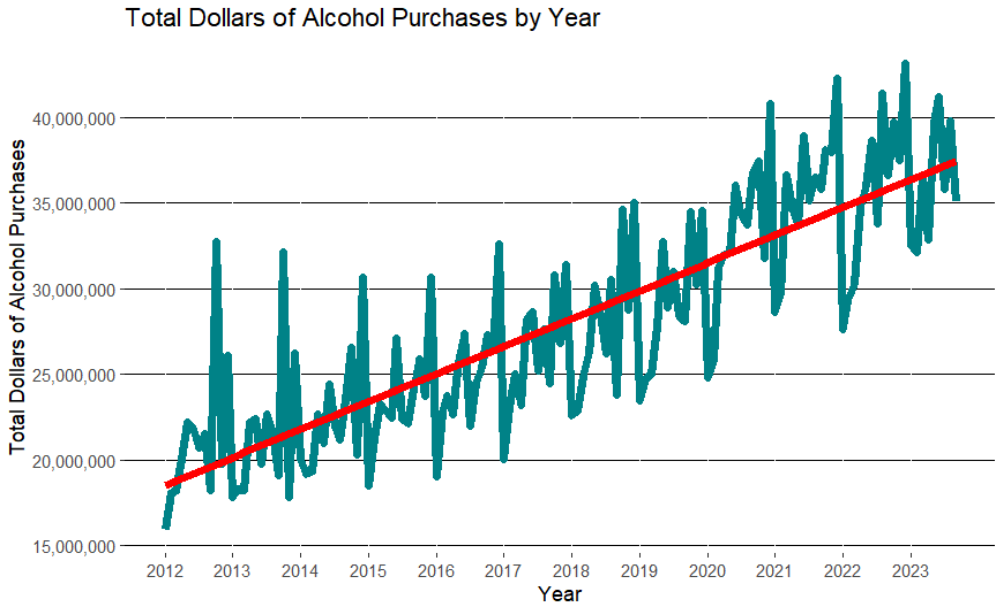

Figure 2: Month-to-month total alcohol purchase amounts in all of Iowa (dollars)

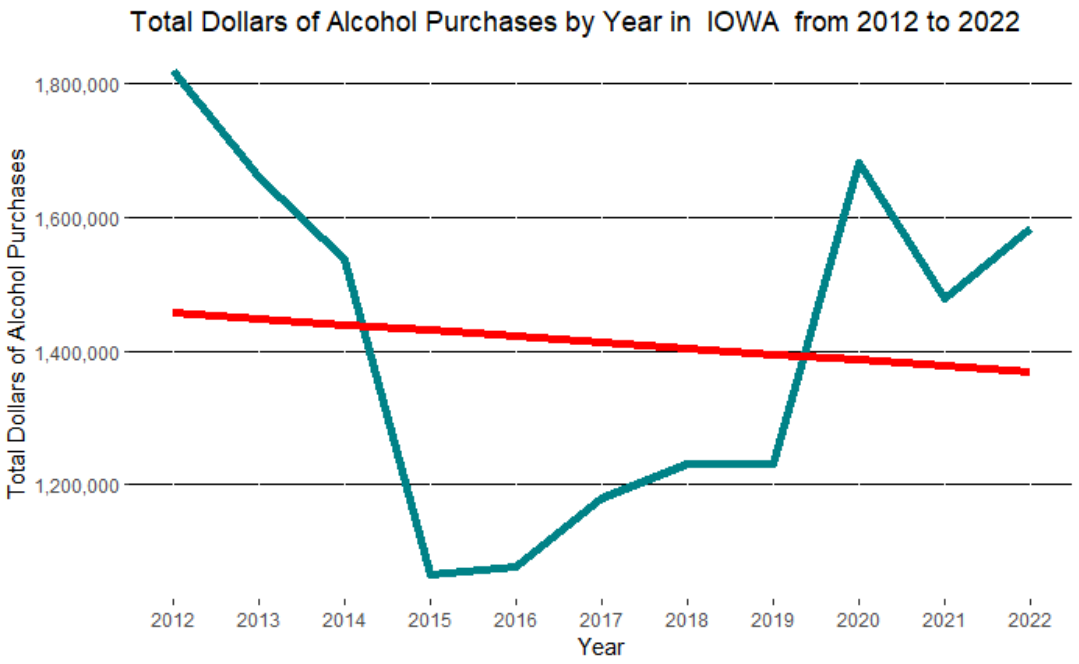

Figure 3: Year-to-year total alcohol purchases in Iowa County by year (dollars)

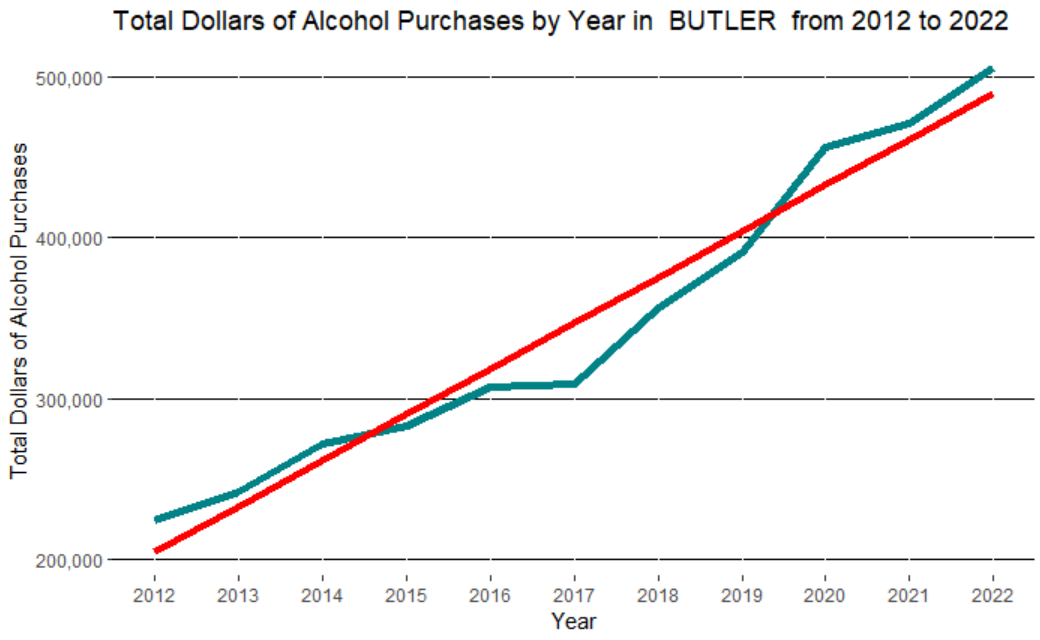

Figure 4: Year-to-year total alcohol purchases in Butler County by year (dollars)